```python
from datetime import datetime, timedelta
from dateutil.parser import parse
import random
import pandas as pd
import numpy as np

import sklearn
from sklearn.metrics import silhouette_score
from sklearn.preprocessing import MinMaxScaler

import matplotlib.pyplot as plt
from matplotlib import cm

from pyclustering.cluster.kmeans import kmeans
from pyclustering.utils.metric import type_metric, distance_metric
from pyclustering.utils.metric \
    import euclidean_distance, manhattan_distance, chebyshev_distance,
    minkowski_distance
from pyclustering.cluster.center_initializer import kmeans_plusplus_initializer


class PycClusteringPeople:
    df_corona = None  # initial loaded data

    added_column_list = []  # Columns added by calculation within class
    target_col_name_list = []  # list for target column names

    sse_list = []  # list for displaying SSE of each cluster
    sil_score_list = []  # list for Silhouette score of clustering result
    centroids_coord_list = []  # list for storing coordinates of centroids

    # member variable for custom distance function
    weight_list = []  # weight values list

    cluster_model_dic = {}  # dictionary for model storing
    scaling_model_dic = {}  # dictionary for model storing

    def __init__(self, file_path, target_col_name_list, base_date):
        self.df_corona = pd.read_csv(file_path)
        self.df_corona["Severity"] = self.compute_severity(base_date, self.df_corona)
        self.added_column_list.append("Severity")

        self.target_col_name_list = target_col_name_list

    def compute_severity(self, base_date, data):
        """
        method to preprocess the data for distance function
        :param base_date: datetime, Base date for calculating elapsed time
        :return: None
        """
        col_num = len(data)  # the number of rows from loaded data
        severity_list = []  # list for storing severity result

        for i in range(col_num):
            # selecting specific column to compute 'severity'
            incur_date_col = data['Incurred Date']
            status = data['Covid Status']
            severity = 0  # default is healthy. 0.

            if status[i] == 'Contacted':  # contacted person?
                # formula for contacted person:
                #   x = 1 - ((today's date) - (infected date)) * 0.05)
                elapsed_days = (base_date - parse(incur_date_col[i]).date()).days
                severity = (1 - (elapsed_days * 0.05)) * 0.5

            elif status[i] == 'Confirmed':  # confirmed person?
                # formula for confirmed person:
```

```
                #   x = (1 - ((today's date) - (infected date)) * 0.05)) / 2
                elapsed_days = (base_date - parse(incur_date_col[i]).date()).days
                severity = 1 - (elapsed_days * 0.05)

            # add the value to the list
            # and rounding to solve floating-point problems
            severity_list.append(round(severity, 4))

        return severity_list

    def display_clustering_result(self,
                                num_cluster,
                                cluster_idx_list,
                                cluster_predicted_list):
        """
        function to display clustering result on console as tabular type
        :param num_cluster: int, the number of cluster
        :param cluster_idx_list: list, cluster index list, ie. [2, 3, 4, 5, 6]
        :param cluster_predicted_list: list, result of clustering
        :return: None
        """

        severity_list = self.df_corona["Severity"].values.tolist()
        age_list = self.df_corona["Age"].values.tolist()

        if type(cluster_predicted_list) != list:
            cluster_predicted_list = cluster_predicted_list.tolist()
        people_num_of_a_cluster_list = []
        avg_age_of_a_cluster_list = []
        avg_severity_of_a_cluster_list = []

        print(f"Number of Clusters: {len(cluster_idx_list)}")

        for cluster_idx in cluster_idx_list:  # 1 cluster
            num_people = cluster_predicted_list.count(cluster_idx)
            id_target_data_tuple_list = []
            target_severity_list = []
            target_age_list = []

            for person_idx in range(len(cluster_predicted_list)):
                if cluster_idx == cluster_predicted_list[person_idx]:
                    target_severity_list.append(severity_list[person_idx])
                    target_age_list.append(age_list[person_idx])
                    id_target_data_tuple_list.append((
                        person_idx + 1,  # [0] of tuple is id
                        age_list[person_idx],  # [1] of tuple is age
                        round(severity_list[person_idx], 2)))  # [2] of tuple is severity

            people_num_of_a_cluster_list.append(num_people)

            print(f"\tCluster {cluster_idx}:")
            print(f"\t\tNumber of People: {num_people}")
            # print(f"\t\t\t{'ID':<4}{'Age':<4}{'Severity Value'}")
            # for person_in_cluster in id_target_data_tuple_list:
            #     print(f"\t\t\t{person_in_cluster[0]:<4}"
            #           f"{person_in_cluster[1]:<4}"
            #           f"{person_in_cluster[2]}")
            print(f"\t\tMinimum of Age values: {min(target_age_list)}")
            print(f"\t\tMaximum of Age values: {max(target_age_list)}")
            print(f"\t\tAverage of Age values: "
                  f"{round(sum(target_age_list) / len(id_target_data_tuple_list), 2)}")
            print(f"\t\tMinimum of Severity values: {min(target_severity_list)}")
            print(f"\t\tMaximum of Severity values: {max(target_severity_list)}")
            print(f"\t\tAverage of Severity values: "
                  f"{round(sum(target_severity_list) / len(id_target_data_tuple_list), 2)}")
            print(f"\t\tThe Coordinates of Centroid:")
            coords = self.centroids_coord_list[num_cluster - 2][cluster_idx]
            print(f"\t\t\tX1 (Severity): {round(coords[0], 2)}")
            print(f"\t\t\tX2 (Age): {round(coords[1], 2)}")
            try:
                avg_age_of_a_cluster_list.append(
                    round(sum(target_age_list) / len(id_target_data_tuple_list), 2))
            except ZeroDivisionError:
                avg_age_of_a_cluster_list.append(0)

            try:
                avg_severity_of_a_cluster_list.append(
                    round(sum(target_severity_list) / len(id_target_data_tuple_list), 2))
            except ZeroDivisionError:
                avg_severity_of_a_cluster_list.append(0)

            print()  # float 1 line
        self.display_summary_table(people_num_of_a_cluster_list,
                                   avg_age_of_a_cluster_list,
                                   avg_severity_of_a_cluster_list)
        print()  # float 1 line

    def display_load_data(self):
        """
        function to display data
        :return: None
        """
        print(f"Total number of People: {len(self.df_corona)}")
        print(f"{'ID':<4}"
              f"{'Age':<4}"
              f"{'Covid Status':<13}"
              f"{'Severity':<9}"
              f"{'Address':<10}")
        for i in range(len(self.df_corona)):
            print(f"{self.df_corona['ID'][i]:<4}"
                  f"{self.df_corona['Age'][i]:<4}"
                  f"{self.df_corona['Covid Status'][i]:<13}"
                  f"{round(self.df_corona['Severity'][i], 3):<9}"
                  f"{self.df_corona['Address'][i].split()[0]:<10}"
                  )
        print()  # float 1 line
        grouped_status = self.df_corona['Severity'].groupby(self.df_corona['Covid Status'])

        print(f"Number of healthy people: {grouped_status.count()['Healthy']}")
        print(f"Number of contacted people: {grouped_status.count()['Contacted']}")
        print(f"Number of confirmed people: {grouped_status.count()['Confirmed']}")

        print(f"Average Severity of contacted people: "
              f"{round(grouped_status.mean()['Contacted'], 2)}")
        print(f"Average Severity of confirmed people: "
              f"{round(grouped_status.mean()['Confirmed'], 2)}")
        print()  # float 1 line

    def display_summary_table(self,
                              people_num_of_a_cluster_list,
                              avg_age_of_cluster_list,
                              avg_severity_of_cluster_list):
        """
        function to display the data as tabular summary
        :param people_num_of_a_cluster_list: list, the number of people in a cluster
        :param avg_age_of_cluster_list: list,
        :param avg_severity_of_cluster_list:
        :return:
        """
        len_id = 17
        len_p_num = 11
        len_age = 13
        len_sev = 15
        len_sum = len_id + len_p_num + len_age + len_sev

        # top row
        print(f"\t{'-' * (len_sum + 11)}")
        print(f"\t{'Cluster ID':>{len_id}} "
              f"| {'# of People':>{len_p_num}} "
              f"| {'Avg. of Ages':>{len_age}} "
              f"| {'Avg. of Severity':>{len_sev}} ")

        # contents of table
        cluster_id = 0
        for people_num, avg_age, avg_sev in zip(people_num_of_a_cluster_list,
                                             avg_age_of_cluster_list,
                                             avg_severity_of_cluster_list):
            print(f"\t{cluster_id:>{len_id}} "
                  f"| {people_num:>{len_p_num}} "
                  f"| {avg_age:>{len_age}} "
                  f"|{avg_sev:>{len_sev}}")
            cluster_id += 1

        print(f"\t{'-' * (len_id + 1)}"
              f"|{'-' * (len_p_num + 2)}"
              f"|{'-' * (len_age + 2)}"
              f"|{'-' * (len_sev + 2)}-")

        # bottom row
        print(f"\t{'Total':^{len_id}} | {sum(people_num_of_a_cluster_list):>{len_p_num}} |")
        print(f"\t{'SSE':^{len_id}} | {round(self.sse_list[len(people_num_of_a_cluster_list) - 2], 2):>{len_p_num}} |")
        print(f"\t{'Silhouette Score':>{len_id}} "
              f"| {round(self.sil_score_list[len(people_num_of_a_cluster_list) - 2], 2):>{len_p_num}} |")
        print(f"\t{'-' * (len_sum + 11)}")

    def draw_graph(self):
        """
        method to draw clustering result
        :return: None
        """
        pass

    def draw_silhouette(self):
        """
        method to draw graph using silhouette scores
        :return: None
        """
        pass

    def draw_elbow_method(self, sse_list):
        """
        method to draw elbow graph using SSE(Sum of Squares Error)
        :param sse_list: list of SSE
        :return: None
        """
        plt.plot(range(2, 10), sse_list, marker='o')
        plt.xlabel("The Number of Cluster")
        plt.ylabel("SSE")
        plt.show()

    def describe_id(self, new_point, cluster_num, cluster_model):
        # TODO: Display the information of reason that why the data in in the cluster.
        #   id is cluster ID. then, this function should explain about the closest cluster.
```

```python
        #    - 경계선인가? 그렇다면 퍼센티지로 나타낼 수 있는가?
        #    - plotting
        # Done
        #    - 가장 가까운 클러스터의 거리와 두번째로 가까운 클러스터의거리
        #    - 어떤 원소들이 여기에 속하는가?
        centroid_coords_list = self.centroids_coord_list[cluster_num-2]

        idx_distance_tuple_list = []
        for idx in range(len(centroid_coords_list)):
            idx_distance_tuple_list.append(
                (idx, self.weighted_euclidean_distance(new_point, centroid_coords_list[idx])))

        print("\tDistance List (Top 3 nearest Clusters)")
        sorted_list = sorted(idx_distance_tuple_list, key=lambda x: x[1])
        closest_cluster_id = sorted_list[0][0]

        print(f"\t\t{'Cluster ID':>11} |{'Distance':>9}")
        for idx_distance_tuple in sorted_list:
            if sorted_list.index(idx_distance_tuple) > 2:
                break
            print(f"\t\t{idx_distance_tuple[0]:>11} |{round(idx_distance_tuple[1], 3):>9.3f}")
        print()  # float 1 line

        feature_values = self.df_corona[self.target_col_name_list]

        # display part
        print(f"\tList of data belonging to the cluster {closest_cluster_id}:")

        closest_cluster_elements_list = cluster_model.get_clusters()[closest_cluster_id]
        for data_id in closest_cluster_elements_list:
            if closest_cluster_elements_list.index(data_id) % 5 == 0:
                print(f"\t\t{str(feature_values.iloc[data_id, :].values.tolist()):<14}", end='')
            elif closest_cluster_elements_list.index(data_id) % 5 == 4:
                print(f"{str(feature_values.iloc[data_id, :].values.tolist()):<14}")
            else:
                if closest_cluster_elements_list.index(data_id) + 1 == len(closest_cluster_elements_list):
                    print(f"{str(feature_values.iloc[data_id, :].values.tolist()):<14}")
                else:
                    print(f"{str(feature_values.iloc[data_id, :].values.tolist()):<14}", end='')
        print()  # float 1 line

    def find_cluster(self, new_person_data, model, base_date=datetime.today().date()):
        # preprocess of input data
        new_person_data["Severity"] = self.compute_severity(base_date, new_person_data)

        target_data = new_person_data[self.target_col_name_list].__deepcopy__()
        # scaling some columns
        scale_col_name = ['Age']
        for col_name in scale_col_name:
            target_data = self.scale_column(target_data, col_name,
                                            using_enrolled_model=True)

        new_point = target_data.loc[:0, tuple(self.target_col_name_list)].values.tolist()
        # predict result: [cluster_id, cluster_id, ... ,]

        return model.predict(new_point)[0], new_point[0], new_person_data

    def get_cluster_model_dic(self):
        """
         function to return cluster_model_dic
         :return: list, cluster_model_dic
         """
        return self.cluster_model_dic

    def get_scaling_model_list(self):
        """
         function to return scaling_model_dic
         :return: list, scaling_model_dic
         """
        return self.scaling_model_dic

    def initialize_random_centroid(self, num_centroid, is_0_1_normalized=True):
        """
         A function that generates a centroid of random coordinates-
         -as many as the number of clusters received.
         :param num_centroid: int, the number of centroids
         :param is_0_1_normalized: boolean, Whether the feature is normalized
         :return: the random coordinates of centroids list
         """
        if is_0_1_normalized:  # are all columns normalized to 0-1?
            # for i in num_centroid:
            return [[random.uniform(0, 1), random.uniform(0, 1)] for _ in range(num_centroid)]

        else:  # there is an unnormalized column
            pass

    def pyc_cluster_kmeans(self,
                           num_cluster,
                           weight_list,
                           distance_function):
        """
         function to cluster data
         :param num_cluster: int, the number of clusters
         :param weight_list: list, weight list of features
         :param distance_function: string, the abbreviation of distance function
         :return: list, clustered result
         """
        self.weight_list = weight_list

        # my_distance_function = lambda p1, p2: p1[0] + p2[0] + 2
        if distance_function == 'eu':
            # metric = distance_metric(type_metric.EUCLIDEAN)
            metric = euclidean_distance
        elif distance_function == 'ma':
            # metric = distance_metric(type_metric.MANHATTAN)
            metric = manhattan_distance
        elif distance_function == 'mi':
            # metric = distance_metric(type_metric.MINKOWSKI)
            metric = minkowski_distance
        elif distance_function == 'c_eu':
            metric = distance_metric(type_metric.USER_DEFINED, func=self.weighted_euclidean_distance)

        target_data = self.df_corona.loc[:, self.target_col_name_list]  # To select required data

        scale_col_name = ['Age']
        for col_name in scale_col_name:
            target_data = self.scale_column(target_data, col_name)
        # if "Age" in self.target_col_name_list:  # scaling only "Age" column
        #     age = target_data.loc[:, "Age"].values.reshape(-1, 1)
        #     scaler = preprocessing.MinMaxScaler()
        #     # data = scaler.fit_transform(data)    # To scale data from 0 to 1
        #     target_data.loc[:, "Age"] = scaler.fit_transform(age)

        # set the number of data and centroids
        self.data_cal_count = num_cluster * len(target_data)
        self.num_of_data = num_cluster * len(target_data)
        self.cent_cal_count = 1
        self.num_of_cent = 1

        # initializing centroids
        # initial_centers = kmeans_plusplus_initializer(target_data, num_cluster).initialize()
        # initial_centers = self.initialize_random_centroid(num_cluster)
        initial_centers = [[i * 0.1, i * 0.1] for i in range(num_cluster)]

        kmeans_instance = kmeans(target_data, initial_centers, metric=metric)
        self.cluster_model_dic[num_cluster] = kmeans_instance

        kmeans_instance.process()
        clustered_list = kmeans_instance.get_clusters()
        clustered_list = self.pyc_result_to_column(clustered_list, len(target_data))

        # add the column
        self.df_corona['Cluster ID'] = clustered_list

        # storing the coordinates of centroids
        self.centroids_coord_list.append(kmeans_instance.get_centers())

        # storing SSE(Sum of Squared Errors)
        self.sse_list.append(kmeans_instance.get_total_wce())

        # strong Silhouette Score
        self.sil_score_list.append(silhouette_score(target_data, clustered_list))

        return clustered_list

    def pyc_result_to_column(self, pyc_cluster_result, people_num):
        """
         function to change shape of Pyclustering to pandas
         :param pyc_cluster_result: nd list, result of clustering using Pyclusterin
         :param people_num: int, the number of people(data)
         :return: list, re-shaped list
         """
        clustered_list = [0 for _ in range(people_num)]

        cluster_id = 0
        for id_list_of_a_cluster in pyc_cluster_result:
            for idx in id_list_of_a_cluster:
                clustered_list[idx] = cluster_id
            cluster_id += 1

        return clustered_list

    def plot_data(self, num_cluster, additional_data=None):
        """
         To plot result
         :return:
         """
        groups = self.df_corona.groupby("Cluster ID")
        fig, ax = plt.subplots()
        for name, group in groups:
            ax.plot(group.Severity, group.Age, marker='o', linestyle="", label=name)
        if additional_data is not None:
            ax.plot(additional_data[0], additional_data[1],
                    marker='*', linestyle="", label='New Data', markersize=15)

        ax.legend(fontsize=12)
        plt.title("Result of Clustering (K=" + str(num_cluster) + ", Weight=" + str(self.weight_list) + ")")
        plt.xlabel("Severity")
        plt.ylabel("Age")
```

```python
        # plt.show()
        file_path = "./Cluster_Result_Plotting_pyc/cluster_result_" + str(num_cluster) + "_" + str(self.weight_list) + ".png"
        if additional_data is not None:
            file_path = file_path[:-4] + '_new_data_plot_.png'
        fig.savefig(file_path, dpi=300)
        plt.close()

    def scale_column(self, data_frame, col_name, feature_range=(0, 1),
                     model_enroll=True,
                     using_enrolled_model=False):
        """
        function to scale data as 0~1
        :param data_frame: pandas dataframe, original data.
        :param col_name: string, column name to scale
        :return: scaled data frame
        """
        origin_data = data_frame.loc[:, col_name].values.reshape(-1, 1)
        if using_enrolled_model:
            model_enroll = False
            scaler = self.scaling_model_dic[col_name]
            data_frame.loc[:, col_name] = scaler.transform(origin_data)
        else:
            scaler = MinMaxScaler(feature_range=feature_range)
            data_frame.loc[:, col_name] = scaler.fit_transform(origin_data)

        if model_enroll:  # storing the scaling model
            self.scaling_model_dic[col_name] = scaler

        return data_frame

    def save_as_csv(self, num_cluster):
        """
        function to save data as .csv file
        :param num_cluster: int, the number of cluster.
        :return: None
        """
        temp_df = self.df_corona.__deepcopy__()

        file_name = f"clustered_corona_data_k={num_cluster}_" \
                    f"{'Severity_Age'}_{''.join(str(self.weight_list))}.csv"
        temp_df.to_csv(file_name, encoding='utf-8-sig')

    def weighted_euclidean_distance(self, point1, point2):
        """
        custom distance function
        :param point1: list, list of feature values or coordinates list of centroid
        :param point2: coordinates list of centroid
        :return: distance between point1 and point2
        """
        distance = 0.0  # distance between point1 and point2

        # when calculating the distance of two coordinates
        if np.shape(point1) == (2,):
            # point 1 is data.
            # point 2 is centroid
            for weight, p1_coord, p2_coord in zip(self.weight_list, point1, point2):
                distance += weight * (p1_coord - p2_coord) ** 2.0

        else:  # when updating Centroid
            # point 1 and 2 are centroid
            for prev_cent, curr_cent in zip(point1, point2):
                for weight, pc_coord, cc_coord in zip(self.weight_list, prev_cent, curr_cent):
                    distance += weight * (pc_coord - cc_coord) ** 2.0

        return distance ** 0.5
```